FERMO: a Dashboard for Streamlined Rationalized Prioritization of Molecular Features from Mass Spectrometry Data

Mitja M. Zdouc[1]*, Lina M. Bayona Maldonado[2], Hannah E. Augustijn[1,2], Sylvia Soldatou[3], Niek de Jonge[1], Marcel Jaspars[3], Gilles P. van Wezel[2], Marnix H. Medema[1]*, Justin J. J. van der Hooft[1,4]*

1 Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands

2 Institute of Biology, Leiden University, Sylviusweg 72, 2333 BE Leiden, The Netherlands

3 Marine Biodiscovery Centre, Department of Chemistry, University of Aberdeen, Old Aberdeen, AB24 3DT, Scotland, United Kingdom

4 Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg 2006, South Africa

* Corresponding authors: Mitja M. Zdouc (mitja.zdouc@wur.nl), Marnix H. Medema (marnix.medema@wur.nl), Justin J. J. van der Hooft (justin.vanderhooft@wur.nl)

ABSTRACT

Small molecules can selectively modulate biological processes and thus generate phenotypic variation. Biological samples are complex matrices, and liquid chromatography tandem mass spectrometry often detects hundreds of molecules, of which only a fraction may be associated with this variation. The challenge therefore lies in the prioritization of the most relevant molecules for further investigation. Tools are needed to effectively contextualize mass spectrometric data with phenotypical and environmental (meta)data. To accelerate this task, we developed FERMO, a dashboard application combining mass spectrometry data with qualitative and quantitative biological observations. FERMO's centralized interface enables users to rapidly inspect data, formulate hypotheses, and prioritize molecules of interest. We demonstrate the applicability of FERMO in a case study on antibiotic activity of bacterial extracts, where we successfully prioritized the bioactive molecule siomycin out of 143 molecular features. We expect that besides natural product discovery, FERMO will find application in a wide range of omics-driven fields.

INTRODUCTION

Small molecules encompass many metabolites, drugs, agrochemicals, and pollutants, which selectively interact with biological systems and influence their functioning. As part of the environment, small molecules can cause phenotypic variations, and their investigation is highly relevant in fields like metabolomics, exposomics, or microbiome research [1, 2]. Liquid chromatography – mass spectrometry (LC-MS) is a commonly used analytical method for the untargeted, qualitative and quantitative analysis of small molecules [3]. LC-MS analysis detects molecules as so-called molecular features, which we here define as a detected ion signal for an eluting molecule, with inherent attributes like a specific mass-to-charge ratio (*m/z*) or retention time. Data-dependent acquisition (DDA) liquid chromatography tandem mass spectrometry analysis (LC-MS/MS) can be used to gather information about the collisional fragmentation of molecular features. Ultimately, the chemical structure of a molecule determines its MS/MS fragmentation spectrum, and structurally similar molecules generally show similar MS/MS fragmentation spectra [4]. The structural information therein encoded can be used for annotation and identification of molecular features [5, 6], molecule fingerprinting [7], and *de novo* predictions of molecular structures [8].

Biological and environmental samples are usually complex matrices, and LC-MS/MS analysis can often detect thousands of molecular features for a single sample. The complex and information-dense nature of LC-MS/MS data generally makes it infeasible to manually identify metabolites that are likely responsible for a phenotype of interest, such as an antimicrobial activity, a medical diagnosis or environmental metadata like altitude or water temperature. For this reason, various computational analysis tools for data reduction [9,10,11,12], organization [13,14,15,16,17], and annotation [18] have been developed. Even after these data pre-processing and reduction steps, hundreds of molecular features can remain in the dataset, all potentially responsible for the observed phenotypic variation. Therefore, the selection of the most relevant molecular features for further investigation requires effective prioritization.

Generally, prioritization is attempted by separating ‘interesting’ or ‘relevant’ from ‘not interesting’ or ‘irrelevant’ molecular features based on a series of attributes (e.g., bioactivity, chemical novelty, diversity, molecular weight, source of sample) [19, 20, 21]. However, the definition and weighting of attributes for selection strongly depend on the respective research question. For example, the definition of chemical novelty may range from ‘a molecule with a novel carbon backbone’ to ‘a novel functional group on a known molecule’. One study may prioritize mainly the presence or absence of a molecular feature, while another may focus on fold-changes between biological groups. This variability in selection principles hampers automation, rendering prioritization still a mostly manual task. Currently available software solutions are typically designed with specific tasks and/or research fields in mind i.e., data quality control [22,23], statistical analysis [24,25], molecular feature annotation [26, 27], or natural product drug discovery [28, 29, 30, 31]. Further, there is a lack of tools that combine and visualize multiple (meta)data types and their relationships in one view.

To bridge this gap, we developed FERMO (**F**ormulation of m**E**trics from **R**eproducible **M**olecular feature **O**bjects), a dashboard application for the prioritization of molecular features relevant for explaining biological observations. FERMO connects LC-MS/MS information with qualitative and quantitative biological meta(data), like sampling location, measured biological

activity, phylogenetic affiliation, or environmental measurements. It allows rapid and reproducible selection of specific molecular features based on a discrete set of attributes. Besides inherent attributes, FERMO calculates custom scores to summarize the relationship of samples and molecular features towards internal and external data. Putative associations to biological data are identified, and presented in the centralized dashboard view at one glance. FERMO is capable of analyzing multiple groups of samples in parallel, and guarantees the reproducibility of analyses by recording processing and filtering parameters. In a case study, we demonstrate how FERMO can rapidly pinpoint molecular features putatively responsible for an observed antibiotic activity. Developed with a focus on user-friendliness and supplied with tutorials and a dedicated Wiki on GitHub, FERMO is aims to improve prioritization in fields like natural product research, metabolomics, environmental, food, and agricultural sciences.

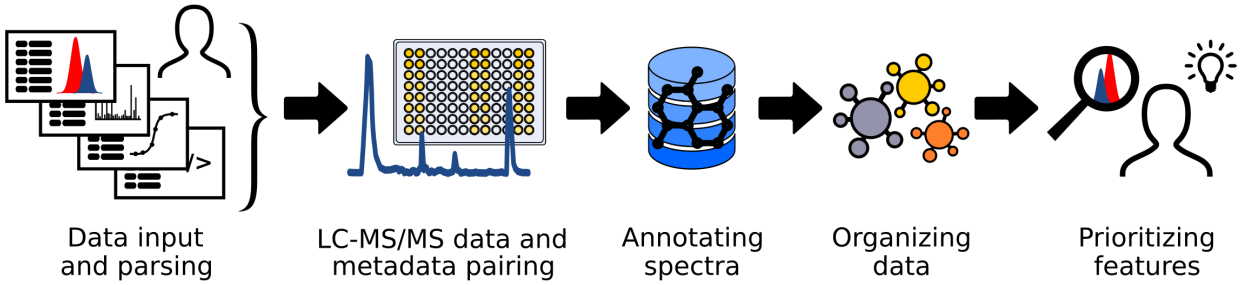

Figure 1: Overview of FERMO processing workflow: LC-MS/MS data, qualitative and quantitative biological (meta)data are parsed and associated with one another. Molecular features are annotated, organized, and presented for prioritization analysis to the user.

RESULTS

<u>FERMO facilitates prioritization *via* custom scores in the dashboard view</u>

FERMO is centered around the dashboard view, an interactive graphical user interface that presents key information about the dataset at a glance (Figure 2). Each sample is associated with a number of descriptive attributes, such as the detected molecular features, calculated scores, group metadata, or quantitative biological data (Figure 2A). Molecular features can be visualized sample-wise in the Sample Chromatogram Overview (Figure 2B) and investigated in detail in the Molecular Feature Information Table (Figure 2C). The Cytoscape Spectral Similarity Networking view (Figure 2D) allows to investigate relationships between molecular features based on MS/MS spectral similarity, and the presence of a molecular feature across samples can be inspected in the Sample Chromatograms view (Figure 2E). Molecular features can be quickly prioritized by combining up to 16 different filters, and the selection can be exported in tabular format for consecutive investigation. A more detailed description of the dashboard design can be found in the Methods section.

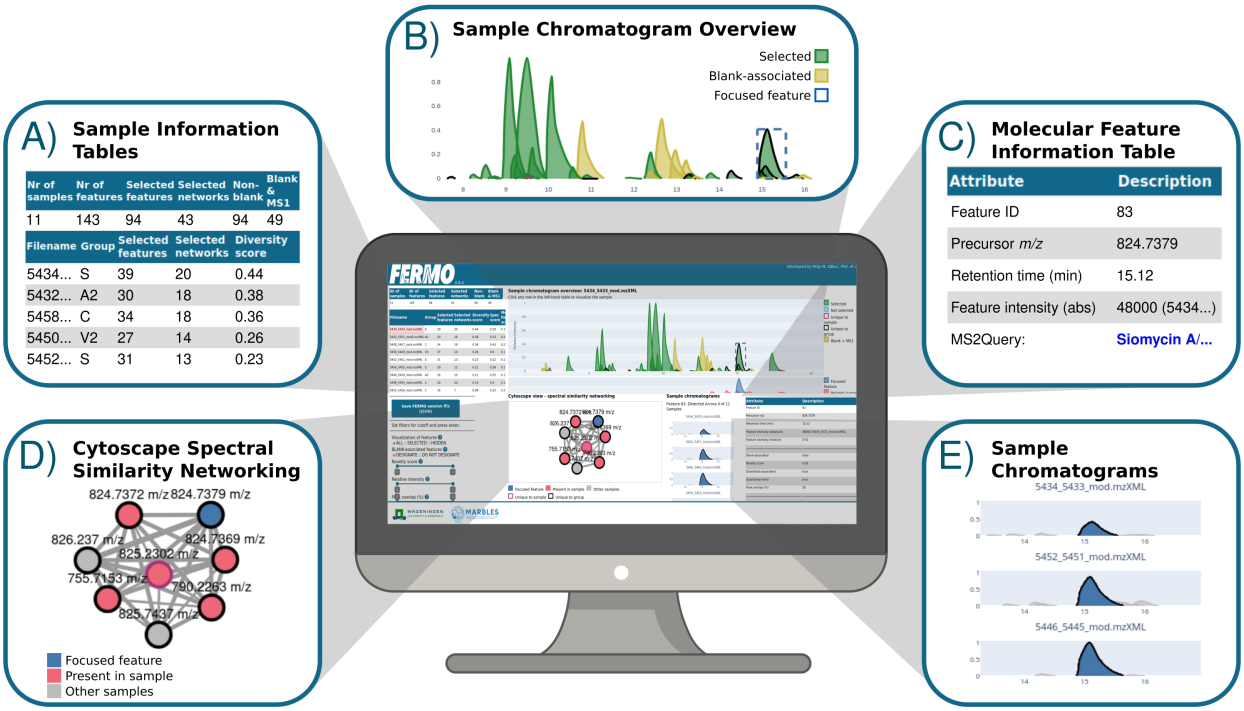

Figure 2: Overview of the FERMO dashboard. A) Sample Information Tables; B) Sample Chromatogram Overview; C) Molecular Feature Information Table; D) Cytoscape Spectral Similarity Networking view; E) Sample Chromatograms.

To facilitate prioritization, FERMO aggregates a number of attributes for each molecular feature and sample. Besides inherent attributes such as precursor *m/z*, retention time, or signal intensity, FERMO calculates several custom, indicative metrics (Table 1). For molecular features, Novelty scores, QuantData scores, and peak collision measures are calculated, while for samples, Diversity scores, Specificity scores, and Mean Novelty scores are provided. The Novelty score estimates the assumed chemical novelty of a molecular feature by comparison to external databases, while the QuantData score designates molecular features that are putatively associated with user-provided quantitative biological data (e.g. biological activity in an assay, days of growth, temperature). The peak collision measure denotes the overlap of co-occurring molecular features and clarifies their ion identities. Diversity and Specificity scores estimate the chemical richness of samples and their associated qualitative biological data (group metadata), respectively, while the Mean Novelty score indicates the assumed overall chemical novelty of a sample. For more details, see the Methods section. These scores allow rapid surveying of LC-MS/MS data by applying filters, retaining only molecular features or samples relevant to the research question. Users are free to restrict the selection as desired, and the filtering parameters are stored to ensure the reproducibility of the analysis.

Table 1: Calculated descriptive attributes for samples and molecular features

Attribute	**Description**
Diversity score	Measure for chemical diversity of sample
Specificity score	Measure for sample-specific chemistry
Mean Novelty score	Summary of putative chemical novelty of molecular features in sample
Peak collision measure	Measure for co-elution of molecular features
Novelty score	Measure for putative novelty of molecular feature
QuantData score	Measure for association of molecular feature to provided quantitative biological data

<u>FERMO associates antibiotic activity to a single family of molecules</u>

To demonstrate the applicability of FERMO, we attempted to pinpoint molecular features responsible for antibiotic activity in a subset of samples from a previously published study on the actinomycete genus *Planomonospora* [32]. We selected a set of samples, based on availability of in-house generated antibiotic activity data against the Gram-positive pathogen *Staphylococcus aureus*. Our set consisted of ten samples: four samples belonged to the S phylogroup, three samples to the C phylogroup, two samples to the A2 phylogroup, and one to the V2 phylogroup. For differentiation between metabolites and growth medium components, we added a sample of the extract of the pure growth medium (referred to as medium blank), leading to a total of eleven samples for the test dataset. Details on the group and biological activity data

can be found in the Supplementary Information. After MZmine3 pre-processing, the resulting data was analyzed by FERMO.

Starting with a total of 143 molecular features extracted by MZmine3, we focused on molecular features that had associated MS/MS information and were not associated with the medium blank (Figure 3A). Association to the medium blank was either direct (molecular features detected both in samples and the blank) or indirect (molecular features not detected in the blank, but co-occurring in spectral similarity networks with directly-associated molecular features). This left 67 molecular features organized in 39 similarity networks. Next, we applied the QuantData score to filter out molecular features that were unlikely to be associated with bioactivity, which left thirteen molecular features in seven spectral similarity networks. The antibiotic activity was only associated with samples belonging to the S group. We assumed that spectral similarity networks containing molecular features from groups other than the S group would be unlikely to be associated with the antibiotic activity. Therefore, we focused on five similarity networks associated only to the S phylogroup, containing ten molecular features. Only three molecular features were detected in all four samples with antibiotic activity (Figure 3B). Inspection of matchms spectral library [33] and MS2Query [34] annotations showed that all three remaining molecular features were annotated as different ions of the antibiotic siomycin A. Siomycin was previously reported to be produced by the actinomyces genus *Planomonospora* and to show growth inhibition against *Staphylococcus aureus* [32,35,36]. Therefore, by data integration and the application of appropriate filtering steps, FERMO was able to rapidly pinpoint the molecular features most likely associated with the observed biological variable (antibiotic activity) and dereplicate the compounds. Since the antibiotic was well-annotated in the study that initially reported the dataset [32], no experimental verification was performed. In other cases, consecutive isolation and structural elucidation of the prioritized compound class are highly recommended.

Increased computational power accelerates data processing

To assess the performance of FERMO on datasets of different sizes and complexity, we conducted exploratory benchmarking. We observed that the length of calculation predominantly depended on the number of molecular features that were investigated, and most computation time was consumed by (i) the annotation of spectra *via* MS2Query [34] and (ii) the calculation of spectral similarity between molecular features *via* matchms. [33] By using a computer with higher computational power, processing time can be decreased (Table 2). We note that for large datasets with several thousand molecular features, calculation time may become unfeasibly long. Work is underway to improve the calculation efficiency and speed by incorporating multi-processing.

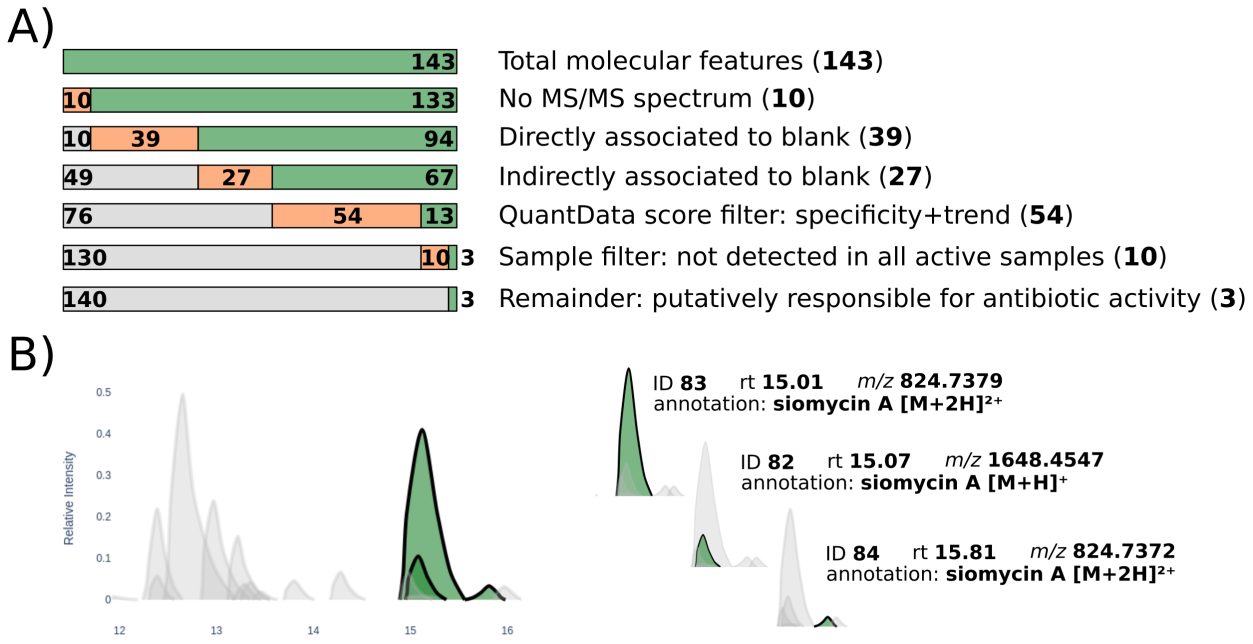

Figure 3: Prioritization procedure in case study. A) application of increasingly restrictive filters leads to the prioritization of three molecular features, likely to be responsible for antibiotic activity. B) inspection of the selected molecular features suggests the antibiotic siomycin A to be responsible for antimicrobial activity.

Table 2: *Exploratory benchmarking results.*

Total molecular features / annotated molecular features	W/o MS2Query (seconds)	With MS2Query (seconds)	Computer specifications
248/154	9.7	670	A
248/154	6.1	308	B
1642/898	489	3792	A
1642/898	280	1859	B
1642/1642	496	6528	A
1642/1642	282	2990	B

A) Processor: Intel® Core™ i5-6200U CPU @ 2.30GHz × 4; Memory: 7.2 GiB; OS: Ubuntu 20.04.5 64bit;
B) Processor: AMD Ryzen™ 5 3600 @ 3.59GHz × 6; Memory 16.0 GiB; OS: Windows 10 Home 64bit;

DISCUSSION

Hypothesis-driven molecular feature prioritization is an essential step in addressing biological observations. FERMO offers a generalized approach for prioritization *via* a rational selection of molecular features based on a number of inherent and calculated attributes. To make FERMO an accessible tool that can be easily included in current workflows, we tried to integrate it with existing software and data ecosystems. In particular, FERMO relies on the program MZmine 3 [9] for LC-MS/MS data pre-processing and peak picking. However, the reliance on a peak table as input format restricts access to the underlying "raw" LC-MS/MS data, and the data quality strongly depends on the parameter settings used in the pre-processing software. While FERMO can not directly mitigate problems arising from suboptimal pre-processing, the FERMO GitHub Wiki provides a detailed tutorial and explanations on pre-processing to inexperienced users. Another, complementary, integrative metabolomics software is available with the recently published tool INVENTA [29]. While FERMO focuses on molecular feature prioritization, INVENTA is aimed at sample prioritization. Due to layout and software design, FERMO is inherently restricted with regard to the amount of data it can efficiently process and visualize, with a soft ceiling of several dozen samples. INVENTA, however, is designed to work with hundreds or thousands of samples at once, and to prioritize mainly on sample-based characteristics, such as estimated novelty based on bibliographic metadata regarding the source material, and estimation of sample chemical diversity by statistical methods. A workflow can be imagined where a large number of samples is first processed by INVENTA, which prioritizes the most promising ones. Consecutively, FERMO can be used to investigate the selected samples in greater detail, to reveal molecular features associated with the biological variable under investigation.

In our case study, we demonstrated how FERMO allows for the rapid prioritization of biologically relevant molecular features. To visualize molecular features in their original, sample-wise organization, FERMO uses a pseudo-chromatogram view similar to a total extracted ion chromatogram (EIC). While an EIC is drawn from a continuous retention time/intensity trace, FERMO constructs pseudo-chromatograms from a discrete number of points (see Methods section). This way of representation can be used to intuitively assess co-eluting molecules and their relationships based on spectral similarity or ion identity. This was shown in the case study, where the bioactivity-associated molecular features were quickly determined to be related, based on their elution pattern, the ion identities, and spectral similarity associations. While pseudo-chromatograms emulate their original counterparts well (SI Figure 2), peak anomalies such as shoulder peaks, fronting, or tailing can lead to misrepresentations. Users are therefore alerted to scrutinize suspicious-looking peaks by consulting the original data using instrument-vendor software, pre-processing software [9, 10, 12], or most conveniently, the recently introduced GNPS Dashboard [22].

To prioritize molecular features, FERMO relies on the calculation of a number of attributes, including the spectral similarity between molecular features, likewise to a number of other tools [37, 38, 39]. In addition, FERMO also allows modification of the underlying similarity calculation, and currently supports Modified Cosine Similarity and MS2DeepScore [15]. FERMO also uses spectral similarity networking to calculate summary scores for sample prioritization. These scores estimate the chemical diversity of a sample, based on the assumption that

spectral similarity networks cluster structurally similar molecular features and can be therefore taken as proxies for chemical classes. While actual chemical class predictions would be desirable, the computational costs associated with more sophisticated approaches such as CANOPUS [40] make routine predictions from hundreds of spectra currently unfeasible. Another prioritization score FERMO uses is the QuantData score, which indicates the putative association of a molecular feature to quantitative biological data, like antibiotic activity in the case study. Several other workflows have been reported to link quantitative biological data to mass spectrometry data, often using statistical tests (e.g. Pearson correlation) [30, 31, 41, 42, 43, 44, 45]. Contrary to these approaches, FERMO does not directly predict the probability of biological activity of a molecular feature. Instead, FERMO determines molecular features that are unlikely to be associated with the quantitative biological measure, based on presence/absence and fold-changes across samples. In the case study, the QuantData score allowed to exclude 69% of candidate molecular features from consideration and to focus on the remainder. The approach is robust towards multiple, unrelated molecular features jointly responsible for an observed biological observation, and does not require sample (pre)fractionation. However, since the QuantData score is based on the intersection of molecular features between samples, only a small number of putative candidates can be excluded if chemically very different samples are compared. For such samples, probability-based approaches may be more suitable.

In conclusion, the FERMO dashboard provides a powerful framework for sample and molecular feature prioritization, based on generalized principles. It pairs intuitive visualization with a range of attribute filters for rapid and streamlined data processing, as demonstrated in the case study. Work is underway to further integrate FERMO with existing tools and data ecosystems, such as allowing input data from other LC-MS/MS data pre-processing tools like OpenMS [12], or XCMS [10], or accepting data formats from advanced prediction tools like CANOPUS [40] or SIRIUS [46]. Another promising avenue is the integration of FERMO and INVENTA in a single workflow, to utilize synergies in the processing of large numbers of samples. Also, the integration of LC-MS/MS data with biological metadata is only the first step in integrative data analysis. The integration of LC-MS/MS metabolomics with other types of omics data, such as metagenomics or transcriptomics, would further strengthen prioritization workflows, as demonstrated with the recently published tools NPOmix [47] or NPLinker [48]. While challenging, such methods promise to further facilitate hypothesis formulation and potentially automate molecular feature prioritization. We anticipate that FERMO's centralized view on LC-MS/MS data analysis will benefit many fields of research in the life sciences.

METHODS

Software design overview

FERMO is a multi-page dashboard app for the prioritization of biologically relevant molecular features from data-dependent acquisition, positive ion mode LC-MS/MS data. FERMO is written in the platform independent Python programming language and may be used on Windows, Mac, and Linux operating systems. The frontend is constructed using the Plotly Dash framework [49], while the backend comprises custom functions for data processing, which utilize a number of public Python packages. A full list of dependencies can be found in the Supplementary Information.

Frontend organization

The FERMO graphical user interface is rendered in a browser application such as Firefox, Google Chrome, Microsoft Edge, or Safari. It is deployed locally, and therefore, no user data is shared *via* the internet. The app is separated into four connected pages: the landing page, the processing mode page, the loading mode page, and the dashboard page. Upon startup, the landing page is shown by default, from where the processing mode page and the loading mode pages can be accessed. The former allows for data processing (see Backend organization section), while the latter can be used to reload a previous FERMO session (Input data formats section). Either page automatically redirects to the dashboard page, where data analysis and molecular feature prioritization take place.

Backend organization

The FERMO backend is a data processing workflow, consisting of functions responsible for data parsing, processing, the calculation of scores, and output data format preparation. FERMO employs both modular design and object-oriented programming principles. Each molecular feature is represented as an individual object with distinct attributes that are modified by the modular processing functions. Therefore, additional functionality can be easily added by defining the respective object attribute and integrating a new function in the workflow. Briefly, the workflow consists of: (i) data parsing; (ii) the determination of association of molecular features to provided metadata (sample grouping data, quantitative biological measurements); (iii) calculation of spectral similarity between molecular features; (iv) annotation of molecular features by matchms spectral library matching and by MS2Query; (v) calculation of peak overlaps and determination of co-occurring adducts and isotopes (resulting from the same analyte); (vi) the calculation of custom scores; (vii) preparation of the visualization. Individual processing steps including the employed parameters are stored in a log file, allowing for reproducible analysis.

Help document and tutorials

On the FERMO GitHub repository, a GitHub Wiki is provided that contains tutorials on input data preparation, data processing, data analysis, and prioritization. In the FERMO app, tool-tips are placed next to parameter input fields, providing impromptu information and link to the respective pages on the FERMO GitHub Wiki. Individual processing functions are documented following the Numpy Documentation Style Guide (https://numpydoc.readthedocs.io/en/latest/format.html).

Deployment and installation

FERMO is open source and can be freely accessed and downloaded *via* its GitHub page (https://github.com/mmzdouc/FERMO/) as a .ZIP-compressed directory, or directly "cloned" using git. FERMO requires Python 3.8 and a number of dependencies (see the Supplementary Information for a detailed list). To install the dependencies, the use of a Python package manager, in particular Conda (https://docs.conda.io/en/latest/), is highly recommended. FERMO comes with a convenient startup script that automatically creates an environment, downloads and installs the dependencies, and starts the program, provided that Conda (either as Anaconda or Miniconda) was installed. If Conda is not available, users may still start the app *via* command line by executing the *app.py* executable. A more thorough description can be found in the README on the FERMO GitHub page.

Versioning

FERMO employs versioning based on Semantic Versioning (https://semver.org/). In particular, this is important for the loading mode, where a previously created analysis (a FERMO session file in the .json format) can be reloaded. In case of patch increment differences (e.g. 0.8.0 and 0.8.1), the session can be loaded nevertheless. In case of minor or major increment differences (e.g. 0.7.0 and 0.8.0), the session file cannot be loaded. If the currently running version of FERMO is incompatible to the one that was used to create the session file, an error message will appear to alert the user.

FERMO data processing

Input data formats

Depending on the mode (processing mode or loading mode), FERMO accepts different input files. In the processing mode, FERMO expects some mandatory, and some optional files. As minimum input, it requires a peak table in the MZmine3 '_quant_full.csv' format, as well as the accompanying .mgf-file. Compatibility with MZmine3 was tested up to version 3.3.0 (SI Table 4). Currently, FERMO expects LC-MS/MS data to be data-dependent acquisition positive ion mode data, which is important for ion identity determination and MS2Query annotation (see the respective sections for further information). Additionally, FERMO accepts group metadata (sample relationships), quantitative (or qualitative) biological data, and a spectral library in the .mgf format. Details on the formatting can be found in the Supplementary Information, or in

the FERMO GitHub Wiki. In the loading mode, FERMO accepts a FERMO session file in the .json format, which allows reloading a previously prepared analysis.

Output data formats

The dashboard view allows for export of several different files: a FERMO session file in the .json format, or peak tables in the .csv-format. The session file contains all information necessary to reload an analysis session in the FERMO loading mode. Besides, this file can be shared with collaborators, containing a precise log of all performed processing steps. Regarding the peak table .csv-file export, the precise format can be adjusted: if desired, exported molecular features can be restricted to the currently selected set and/or sample. Of note, the peak table export automatically creates a second file in the .json format, containing a log of all performed processing steps, for the convenience of the user.

Input testing, data parsing

Upon loading, input data files are parsed and tested for correct formatting. A status message informs the user of the outcome of the test. Once (mandatory) files are loaded, processing can be initiated. Molecular feature objects are initiated, and the association between LC-MS/MS data and metadata is determined. If sample grouping metadata was provided, samples and molecular features are organized after their group affiliation, which is of particular interest for blank-associated molecular features. Samples without attributed grouping information are automatically organized in the “GENERAL” wildcard group.

Calculating similarity networks

During data organization, FERMO calculates the spectral similarity between MS/MS fragmentation spectra of molecular features. Since fragmentation spectra ultimately depend on the chemical structure of the molecule and similar molecules usually yield similar fragmentation spectra, spectral similarity can be taken as a proxy for chemical similarity. FERMO allows to change the similarity calculation algorithm, and allows switching between Modified Cosine similarity [33] and MS2DeepScore [15]. Further parameters, like the desired spectrum similarity score cutoff, can be set on the processing mode page.

Annotating the identity of molecular features

The putative identity of molecular features can be annotated in two ways: first, by spectral library matching against a user-provided library using the Modified Cosine similarity algorithm from the matchms package [33], and second, by matching against an pre-calculated embeddings library of over 300.000 mass spectra, using the MS2Query algorithm [34]. The former is intended to be used with a targeted library of molecules that are suspected to be present in the samples, while the latter employs a generalized library and also detects putative analogs of known molecules. Since MS2Query is a computationally expensive algorithm, it can be restricted to annotate only a subset of molecular features (based on relative intensity and/or

exclude blank-associated ones) or switched off completely. Of note, the MS2Query annotation workflow in FERMO currently only supports positive ion mode LC-MS/MS data.

Calculating peak overlaps and ion identities

In LC-MS, molecular features with distinct *m/z* ratios but overlapping retention time windows can either correspond to different co-eluting molecules, or result from different ion species of the same molecule (e.g., $[M+H]^+$, $[2M+H]^+$, $[M+2H]^{2+}$, $[M+Na]^+$, …). FERMO attempts to identify and annotate ion species resulting from the same molecule. Of note, only positive ion mode is currently supported. First, peak overlaps are determined: for two peaks A and B, let S_A and S_B be the retention time at the start of the peaks, and E_A and E_B the retention time at the end of the peaks (SI Figure 1). Peaks do not overlap if either $E_A < S_B$ or $E_B < S_A$ are true. For each overlapping pair of molecular features, FERMO tests if the difference between their *m/z* ratios correspond to one of 15 different ion and/or isotope adducts (see SI Table 2), assuming that one of the two is the $[M+H]^+$ ion. If a match was found, the molecular features are putatively annotated as related ion species, and their overlap is not registered as co-elution, since they originate from the same molecule. These adduct annotations can be queried *via* the "Adduct/isotope search" filter in the dashboard view, for example to identify iron adducts (suggesting siderophores). If overlapping molecular features cannot be recognized as related ion species, they are assumed to stem from different molecules, and their peak overlap is registered as co-elution. For each molecular feature, the fraction of the peak duration that is not overlapping with other molecular features is calculated, and can be queried *via* the "Peak overlap" filter on the dashboard (a value from 0-1). This allows users to quickly identify molecular features with low or minimal overlap that could be promising candidates for chromatographic isolation.

Calculating Novelty and Mean Novelty score

The Novelty score (a value from 0-1) indicates the putative chemical novelty of the molecular feature, assessed by comparison against external data. It takes into account results from both matching against a user-provided spectral library, and from MS2Query matching against pre-calculated spectral embeddings. The Novelty score is flexible in the sense that it still leads to an output even if one of the annotation methods was not performed (e.g., when no mass spectral library was provided). High-certainty annotations lead to a low Novelty score, while low certainty annotations lead to a higher score. The Novelty score N is calculated as follows: for a molecular feature *f*, let S_L (0-1) be the best library match score, S_M (0-1) the best MS2Query match score, and S_{NN} (0-1) the reciprocal of the number of different NPClassifier/Classyfire superclass annotations of nearest neighbors of *f* in its spectral similarity network (provided by MS2Query annotation). If any of S_L or S_M are higher than 0.95, *f* is assumed to be reliably annotated, and N is returned as 1 - S_L or 1 - S_M, depending on which of S_L, S_M is higher. Else, any of S_L, S_M, S_{NN} greater than 0 are summed, divided by the number of summed elements, and is returned as 1 - sum. S_{NN} is an estimate for the similarity of annotation by MS2Query. Since the next neighbors in the spectral similarity network are hypothesized to be related, they should be also annotated as the same chemical class. If so, it strengthens the annotation certainty and therefore leads to

a lower Novelty score. Consequently, the Mean Novelty score is the mean of Novelty scores of all molecular features per sample, excluding blank-associated molecular features.

Calculating the QuantData score

The QuantData score takes into account association of molecular features to user-provided quantitative biological data. This is often biological activity data, but can be any quantitative biological measurement. For each molecular feature *f*, its presence across all samples is considered. If *f* is only detected in samples associated with the quantitative biological data, it is considered to be putatively associated with this measurement. In cases where *f* is detected in both types of samples, the fold-difference between the lowest intensity of *f* across associated samples and the highest intensity of *f* across samples not associated to the measurement is compared. If this fold-difference is higher than a user-provided QuantData factor (by default, 10), *f* is still putatively associated with the measurement. This is a pragmatic precaution against excluding molecular features that might be present in concentrations too low for a sample to be associated with the quantitative biological data (e.g. subinhibitory concentration), but high enough to be detected by the mass spectrometer. Besides this sample-specificity, also the correlation between molecular feature intensity and the quantitative biological variable (the trend) is assessed. On the dashboard page, users can decide if they want to only consider specificity or both specificity and trend.

Calculating Diversity and Specificity scores

The Diversity and the Specificity scores indicate the chemical diversity that is contained by a sample, compared against all samples in the dataset (i.e., internally). They are intended to be used as consideration in sample prioritization and are based on the assumption that molecular features in a spectral similarity network belong to the same chemical class. Therefore, each network represents a proportion of the chemical diversity of the dataset, and all networks combined represent the total chemical diversity. The Diversity score is calculated by dividing the number of spectral similarity networks associated to a single sample with the total number of similarity networks. A sample with a high Diversity score should therefore contain more chemical diversity than a sample with a low Diversity score. Similarly, the Specificity score indicates the unique chemistry of the group the sample belongs to. Here, the number of networks in the sample, subtracted by the networks also detected in other groups, is divided by the number of networks detected in the sample. In other words, the Specificity score is a fraction of the Diversity score, indicating the proportion of ‘unique’ networks per group. High Diversity and Specificity scores are beneficial parameters in sample prioritization, since a sample with more diversified chemistry might provide multiple interesting compound classes.

Constructing pseudo-chromatograms of molecular features

FERMO represents molecular features as pseudo-chromatograms. Contrary to normal extracted ion chromatograms (EICs) which consist of a continuous retention time/intensity trace for each *m/z* ratio, pseudo-chromatograms are constructed from a discrete number of data points: for each molecular feature, the trace is constructed from the retention times at the beginning and the end of the peak, the molecular feature width at half maximum intensity, and the retention time at the apex of the peak. This approach shows good correspondence to the original EICs (SI Figure 2).

FERMO visualization and analysis

The FERMO dashboard consists of six elements (Figure 2): i) the Sample Information Tables; ii) the Sample Chromatogram Overview; iii) the Molecular Feature Information Table; iv) the Sample Chromatograms; (v) the Cytoscape Spectral Similarity Networking view; (vi) the Filter and Export Panel.

Sample Information Tables

The Sample Information Tables provide an overview of the samples included in the analysis. They contain descriptive statistics, such as the number of selected molecular features and networks, and summarize the calculated sample scores. The currently displayed sample in the Sample Chromatogram Overview can be changed by clicking on one of the rows.

Sample Chromatogram Overview

The Sample Chromatogram Overview shows a pseudo-chromatogram of the selected sample. The top panel shows an overview of the molecular features, color-coded after their attributes. Green indicates that the molecular feature is currently under selection, cyan that it is not selected. Blank-associated molecular features are indicated in yellow. A click on one of the molecular features focuses it and activates the bottom panel of the sample chromatogram view, which indicates related molecular features in the same sample. Furthermore, additional information about the focused molecular feature is displayed in the Molecular Feature Information Table, the Sample Chromatograms, and the Cytoscape Spectral Similarity Networking view. Display of molecular features can be restricted only to the currently selected ones by adjusting the “Visualization of features” toggle.

Molecular Feature Information Table

The Molecular Feature Information Table provides detailed information on the focused molecular feature. This includes general attributes like *m/z* ratio or average retention time, calculated scores like Novelty or QuantData scores, annotations from library matching and MS2Query, and information about the associated spectral similarity network.

Sample Chromatograms

The Sample Chromatograms visualize the presence of the focused molecular feature across samples, and show the spatial context respective peak co-elution across samples. This view allows to quickly identify the most promising sample (with the lowest number of co-eluting molecules) for chromatographic isolation of the molecular feature of interest.

Cytoscape Spectral Similarity Networking view

The Cytoscape Spectral Similarity Networking view shows the spectral similarity network the selected molecular feature is associated with, using a Cytoscape-based plugin [50,51]. In the network, each node indicates a molecular feature, which is color-coded after its attributes. Nodes and edges can be clicked to display more information in a separate table below the Cytoscape view.

Filter and Export Panel

In the Filter and Export Panel, users can select a specific subset of molecular features, using up to 16 different filters (SI Table 5). Molecular features can be filtered for their scores, their association to spectral similarity networks or groups. Of note, the “Adduct/isotope search”, the “Annotation search”, the “Group filter (features)” and the “Group filter (networks)” use POSIX (Portable Operating System Interface for uniX) extended regular expressions. Furthermore, peak tables and molecular feature objects can be exported. Also, a so-called FERMO session file can be saved, which allows for later reloading of the session.

Data collection and preparation for the case study

Antimicrobial activity testing:

Testing of antimicrobial activity of strains was performed by agar diffusion assay in Naicons Srl. laboratories (Milan, Italy). Briefly, *Staphylococcus aureus* L100 (Naicons Srl. Milan, Italy) was inoculated in 30 mL of liquefied hand warm Müller-Hinton Agar, with a final concentration of $1x10^5$ CFU x mL^{-1}. After solidification, 10 µL of each sample (dissolved in a 50 % methanol in water solution (v/v)) was applied as a distinct droplet and left to dry. As positive control, 5 µL of a solution of Apramycin 5 mg x mL^{-1} in sterile water was applied in a likewise fashion. As negative control, 10 µL of a 50 % methanol in water solution (v/v) was applied. Plates were incubated at 37 °C overnight and the diameter of the growth inhibition zones was measured in millimeters. The positive controls showed an average growth inhibition zone of 22 mm, while no growth inhibition was observed for the negative control. The results can be found in SI Table 3 in the Supporting Information of this publication.

Analytical conditions, data retrieval and subsetting

In this study, no new LC-MS data was generated, and we only re-analyzed previously published data [32]. Samples were selected based on availability of antimicrobial activity data. In total, 10 samples and one pure medium extract (referred to as medium blank) were used. Since samples were already in the .mzXML format, no further processing was performed, and samples were directly imported into MZmine 3.2.8.

MZmine3 pre-processing parameters

For processing, MZmine version 3.2.8 was used. After data import, the following workflow was employed: (A) MassDetection = retention time, auto; MS 1 noise level, 1E3; MS 2 noise level, 2E1. (B) ADAP chromatogram builder [52] = retention time, auto; MS-level, 1; min group size in no. of scans, 8; group intensity threshold, 5E2; min highest intensity, 1E3; *m/z* tolerance, 20 ppm. (C) Chromatogram deconvolution = local minimum (feature) resolver; Chromatographic threshold, 85%; Minimum search range RT (absolute), 0.05; Minimum relative height, 0%, Minimum absolute height, 5.0E3; Min ratio of peak top/edge, 1.7; Peak duration range (min) 0.15-1.00; Min # of data points, 5; *m/z* range MS 2 pairing, 0.02; RT range MS 2 pairing, 0.4 min. (D) 13C Isotope Filter = *m/z* tolerance, 20 ppm; RT tolerance, 0.2 min; monotonic shape, no; maximum charge, 2; representative isotope, lowest *m/z.* (E) RANSAC peak alignment = *m/z* tolerance, 20 ppm; RT tolerance, 0.7 min; RT tolerance after correction, 0.35 min; RANSAC iterations, 100 000; minimum number of points, 50%; threshold value, 0.5; linear model, no; require same charge state, no. (F) Duplicate peak filter = filter mode, new average; *m/z* tolerance, 0.01 *m/z* or 20 ppm; RT tolerance, 0.1 min. Molecular features with a retention time <1.5 min or >20 min were excluded to remove the initial solvent peak and the column wash phase, respectively. The resulting molecular feature list contained 143 entries and was exported *via* 'Feature list methods' → 'Export feature list' → 'GNPS feature based molecular networking' with settings = merge MS/MS, no; Filter rows, ALL; Feature intensity, Peak height; CSV export, ALL.

FERMO parameters and analysis settings

Processing was performed using FERMO version 0.8.1. Briefly, the peak table and .mgf file created by MZmine 3.2.8 processing, as well as the group information metadata table, the biological activity table, and an in-house spectral library were loaded on the Processing mode page. The format of the quantitative biological data (=biological activity data) was specified as percentage-like (i.e. highest value is highest activity), since growth inhibition was measured in millimeters, and a wider growth inhibition radius signifies a more potent antimicrobial inhibition. Parameters were set as follows: Mass deviation, 20 (ppm); Min fragments per MS^2 spectrum, 8; QuantData factor, 10; Blank factor, 10; Blank factor, 10; Relative intensity filter, 0-1; MS2Query, ON; MS2Query relative intensity filter, 0-1; MS2Query - annotate features from blanks, OFF; Spectral similarity networking algorithm, 'Modified cosine'; Fragment similarity tolerance, 0.1; Spectrum similarity score cutoff, 0.8; Max spectral links, 10; Min matched peaks, 8. Prioritization was performed with the following settings: QuantData-associated, 'SPEC+TREND'; Group filter (networks), ^S$ (signifying 'only group S'); Number samples filter, 4 (minimum).

DATA AVAILABILITY

Mass spectrometry data for the case study was taken from **MSV000085376** (https://doi.org/doi:10.25345/C5412V) [32]. FERMO input files used in the case study (SI Table 6) can be found in the 'example_data' folder in the FERMO GitHub repository (https://github.com/mmzdouc/FERMO/).

CODE AVAILABILITY

FERMO is open source and is freely available on its GitHub page (https://github.com/mmzdouc/FERMO/), under the permissive MIT license. The FERMO Wiki is available on GitHub (https://github.com/mmzdouc/FERMO/wiki/).

REFERENCES

[1] Niessen, W. MA. Interpretation of MS-MS mass spectra of drugs and pesticides. John Wiley & Sons (2017).

[2] Deblonde, T. et al. Emerging pollutants in wastewater: a review of the literature. *J. Hyg. Environ. Health.* **214**, 442-448 (2011).

[3] Wolfender, JL. et al. Accelerating metabolite identification in natural product research: toward an ideal combination of liquid chromatography-high-resolution tandem mass spectrometry and NMR profiling, in silico databases, and chemometrics. *Anal. Chem.* **91**, 704-742 (2018).

[4] Quinn, R.A. et al. Molecular networking as a drug discovery, drug metabolism, and precision medicine strategy. *Trends Pharmacol. Sci.* **38**, 143-154 (2017).

[5] Kind, T. et al. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom. Rev.* **37**, 513-532 (2018).

[6] Beniddir, M.A. et al. Advances in decomposing complex metabolite mixtures using substructure and network-based computational metabolomics approaches. *Nat. Prod. Rep.* **38**, 1967-1993 (2021).

[7] Ludwig, M. et al. Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints. *Bioinformatics,* **34**, i333–i340 (2018).

[8] Stravs, M.A. et al. MSNovelist: De novo structure generation from mass spectra. *Nat. Methods* **19**, 865-870 (2022).

[9] Pluskal, T. et al. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, **11**, 395 (2010).

[10] Smith, C.A. et al. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **78**, 779-787 (2006).

[11] Tautenhahn, R. et al. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Anal. Chem.* **84**, 5035–5039 (2012).

[12] Sturm, M. et al. OpenMS – An open-source software framework for mass spectrometry. *BMC Bioinformatics*, **9**, 163 (2008).

[13] Yang, J.Y. et al. Molecular Networking as a Dereplication Strategy. *J. Nat. Prod.* **76**, 1686–1699 (2013).

[14] Nguyen, D.D. et al. MS/MS networking guided analysis of molecule and gene cluster families. *Proc. Natl. Acad. Sci. U.S.A*. **110**, E2611-E2620 (2013).

[15] Huber F. et al. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J. Cheminformatics*. **13**, 84 (2021).

[16] Huber, F. et al. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput. Biol.* **17**, e1008724 (2020).

[17] Watrous, J. et al. Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U.S.A*. **109**, E1743-E1752 (2012).

[18] Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299-302 (2019).

[19] Donadio, S. et al. Approaches to discovering novel antibacterial and antifungal agents. Methods Enzymol. **458**, 3-28 (2009).

[20] Tabudravu, J.N. et al. LC-HRMS-Database Screening Metrics for Rapid Prioritization of Samples to Accelerate the Discovery of Structurally New Natural Products. *J. Nat. Prod.* **82**, 211–220 (2019).

[21] Zdouc, M.M. et al. A biaryl-linked tripeptide from Planomonospora reveals a widespread class of minimal RiPP gene clusters. *Cell Chem. Biol.* **28**, 733-739 (2021).

[22] Petras, D. et al. GNPS Dashboard: collaborative exploration of mass spectrometry data in the web browser. *Nat. Methods* **19,** 134–136 (2022).

[23] Letourneau, D.R. et al. Constellation: An Open-Source Web Application for Unsupervised Systematic Trend Detection in High-Resolution Mass Spectrometry Data. *J. Am. Soc. Mass Spectrom.* **33**, 382–389 (2022).

[24] Xia, J. et al. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res*. **37**, W652–W660 (2009).

[25] Sample, R. et al. MPACT: An Advanced Informatics Tool for Metabolomics and Data Visualization of Specialized Metabolites from Complex Microbial Samples. ChemRXiv https://doi.org/10.26434/chemrxiv-2022-r0xbx (2022).

[26] Koelmel, J. et al. Interactive Software for Visualization of Non-Targeted Mass Spectrometry Data – FluoroMatch Visualizer. *ChemRXiv* https://doi.org/10.26434/chemrxiv-2022-p5l50 (2022).

[27] McEachran, D.A. et al. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. *Anal. Bioanal. Chem.* **409**, 1729–1735 (2017).

[28] Pharm T.H. et al. Species Prioritization Based on Spectral Dissimilarity: A Case Study of Polyporoid Fungal Species. *J. Nat. Prod.* **84**, 298-309 (2021).

[29] Quiros-Guerrero, LM. et al. Inventa: A computational tool to discover structural novelty in natural extracts libraries. *Front. Mol. Biosci.* 9:1028334 (2022).

[30] Olivon, F. et al. Bioactive Natural Products Prioritization Using Massive Multi-informational Molecular Networks. *ACS Chem. Biol.* **12**, 2644–2651 (2017).

[31] Sanghoon, L. et al. NP Analyst: An Open Online Platform for Compound Activity Mapping. *ACS Cent. Sci.* **8**, 223–234 (2022).

[32] Zdouc M.M. et al. Planomonospora: A metabolomics perspective on an underexplored Actinobacteria genus. *J. Nat. Prod.* **84**, 204-219 (2021).

[33] Huber, F. et al. matchms - processing and similarity evaluation of mass spectrometry data. *J. Open Source Softw.* **5**, 2411 (2020).

[34] De Jonge, N. et al. MS2Query: Reliable and Scalable MS2 Mass Spectral-based Analogue Search. *bioRxiv* https://doi.org/10.1101/2022.07.22.501125 (2022).

[35] Ebata, M. Studies on siomycin. I Physicochemical properties of siomycins A, B and C. *J. Antibiot.* **22**, 364-368 (1969).

[36] Thiemann, J.E. et al. Antibiotic Production by New Form-Genera of the Actinomycetales. I Sporangiomycin, an Antibacterial Agent Isolated from Planomonospora Parontospora var. Antibiotica var Nov. *J. Antibiot.* **21**, 525-531 (1968).

[37] Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotech.* **34**, 828–837 (2016).

[38] Olivon, F. et al. MetGem Software for the Generation of Molecular Networks Based on the t-SNE Algorithm. *Anal. Chem.* **90**, 13900–13908 (2018).

[39] Nothias, LF. et al. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17**, 905–908 (2020).

[40] Dührkop, K. et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* **39**, 462-471 (2021).

[41] Nothias, LF. et al. Bioactivity-Based Molecular Networking for the Discovery of Drug Leads in Natural Product Bioassay-Guided Fractionation. *J. Nat. Prod.* **81**, 758–767 (2018).

[42] Aligiannis, N. et al. Heterocovariance Based Metabolomics as a Powerful Tool Accelerating Bioactive Natural Product Identification. *ChemistrySelect,* **1**, 2531-2535 (2016).

[43] Bertrand, S. et al. Statistical Correlations between HPLC Activity-Based Profiling Results and NMR/MS Microfraction Data to Deconvolute Bioactive Compounds in Mixtures. *Molecules*. **21**, 259 (2016).

[44] Kurita, K.L. et al. Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. *Proc. Natl. Acad. Sci. U.S.A*. **112,** 11999-12004 (2015).

[45] Mladic, M. et al. At-line nanofractionation with parallel mass spectrometry and bioactivity assessment for the rapid screening of thrombin and factor Xa inhibitors in snake venoms. *Toxicon* **110**, 79-89 (2016).

[46] Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* ***16****,* 299-302 (2019).

[47] Leão, T. F. et al. NPOmix: a machine learning classifier to connect mass spectrometry fragmentation data to biosynthetic gene clusters. *Proc. Natl. Acad. Sci. U.S.A. Nexus* (2022).

[48] Hjörleifsson, E. G. et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. *PLoS Comput. Biol.* **17,** e1008920 (2021).

[49] Shammamah, H. Visualization of Bioinformatics Data with Dash Bio. https://doi.org/10.25080/Majora-7ddc1dd1-012 (2019).

[50] Shannon, P. et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498-2504 (2003).

[51] Franz, M. et al. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* **32**, 309–311 (2015).

[52] Myers, O.D. et al. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Anal. Chem.* **89**, 8696-8703 (2017).

Acknowledgements

The authors thank Margherita Sosio and Stefano Donadio (both Naicons Srl., Milano, Italy) for the use of antibiotic activity data gathered at Naicons Srl. laboratories in the present case study. The authors also thank Luis Manuel Quirós-Guerrero, Louis-Félix Nothias, Adriano Rutz, and Jean-Luc Wolfender for valuable discussion. Further, the authors thank the participants of the 2022 Dagstuhl Conference with the title: *"Computational Metabolomics: From Spectra to Knowledge"* for valuable inspiration. Furthermore, the authors are thankful to the beta version testers for their effort and feedback (in alphabetical order): Marianna Iorio, Matteo Simone, Soliman Khatib, and Sonia Maffioli.

Author Information

Corresponding Authors

Mitja M. Zdouc - Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands; ORCID: https://orcid.org/0000-0001-6534-6609; Email: mitja.zdouc@wur.nl

Marnix H. Medema - Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands; ORCID: https://orcid.org/0000-0002-2191-2821; Email: marnix.medema@wur.nl

Justin J. J. van der Hooft - Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands; Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg 2006, South Africa; ORCID: https://orcid.org/0000-0002-9340-5511; Email: justin.vanderhooft@wur.nl

Authors

Lina M. Bayona Maldonado - Institute of Biology, Leiden University, Sylviusweg 72, 2333 BE Leiden, The Netherlands; ORCID: https://orcid.org/0000-0002-5026-3621

Hannah E. Augustijn - Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands; Institute of Biology, Leiden University, Sylviusweg 72, 2333 BE Leiden, The Netherlands; ORCID: https://orcid.org/0000-0002-1862-6699

Sylvia Soldatou - Marine Biodiscovery Centre, Department of Chemistry, University of Aberdeen, Old Aberdeen, AB24 3DT, Scotland, United Kingdom; ORCID: https://orcid.org/0000-0002-3868-102X

Niek F. de Jonge - Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands; ORCID: https://orcid.org/0000-0002-3054-6210

Marcel Jaspars - Marine Biodiscovery Centre, Department of Chemistry, University of Aberdeen, Old Aberdeen, AB24 3DT, Scotland, United Kingdom; ORCID: https://orcid.org/0000-0002-2426-6028

Gilles P. van Wezel - Institute of Biology, Leiden University, Sylviusweg 72, 2333 BE Leiden, The Netherlands; ORCID: https://orcid.org/0000-0003-0341-1561

Contributions

M.M.Z. developed the methods, packages, and wrote the help documents, wiki, and tutorials. M.M.Z. and L.M.B.M. prepared and analyzed case study data. M.H.M. and J.J.J.vdH. supervised development and implementation of the software. M.M.Z. and H.E.A. designed the frontend of the software. M.M.Z. and H.E.A. prepared the figures. L.M.B.M., H.E.A., S.S., and N.dJ. tested the software. H.E.A. and N.dJ. conducted code review. L.M.B.M., H.E.A., S.S., N.dJ., S.S., M.J., G.P.vW., M.H.M, and J.J.J.vdH. provided valuable input on software functionality and improvement. M.M.Z. wrote the manuscript and M.H.M., and J.J.J.vdH. improved the manuscript. All authors contributed to and approved of the final manuscript.

Corresponding authors

Correspondence to Mitja M. Zdouc and Marnix H. Medema and Justin J.J. van der Hooft.

Ethics declarations

Competing interests

J.J.J.vdH. is a member of the Scientific Advisory Board of Naicons Srl., Milano, Italy. M.H.M. is a member of the scientific advisory board of Hexagon Bio and co-founder of Design Pharmaceuticals. The other authors declare no competing interests.

Funding

This work was funded by the European Union Horizon 2020 project MARBLES [101000392].